Airbnb Data Analysis

Ram Kumar

2024-08-15

load data

```
library(lubridate)
airbnb <- read_csv("D:/Airbnb_project/NYC_Airbnb_Open_Data.csv")
airbnb <- airbnb %>%
  mutate(
    price = as.numeric(price),
    reviews_per_month = replace_na(reviews_per_month, 0),
    last_review = dmy(last_review)
  ) %>%
  filter(!is.na(price) & price > 0)
head(airbnb)
```

```
## # A tibble: 6 x 16
##      id name        host_id host_name neighbourhood_group neighbourhood latitude
##   <dbl> <chr>         <dbl> <chr>     <chr>               <chr>            <dbl>
## 1  2539 Clean & qu~    2787 John      Brooklyn            Kensington        40.6
## 2  2595 Skylit Mid~    2845 Jennifer  Manhattan           Midtown           40.8
## 3  3647 THE VILLAG~    4632 Elisabeth Manhattan           Harlem            40.8
## 4  3831 Cozy Entir~    4869 LisaRoxa~ Brooklyn            Clinton Hill      40.7
## 5  5022 Entire Apt~    7192 Laura     Manhattan           East Harlem       40.8
## 6  5099 Large Cozy~    7322 Chris     Manhattan           Murray Hill       40.7
## # i 9 more variables: longitude <dbl>, room_type <chr>, price <dbl>,
## #   minimum_nights <dbl>, number_of_reviews <dbl>, last_review <date>,
## #   reviews_per_month <dbl>, calculated_host_listings_count <dbl>,
## #   availability_365 <dbl>
```

summary analysis

```
summary(airbnb[, c("price", "minimum_nights", "number_of_reviews", "availability_365")])
```

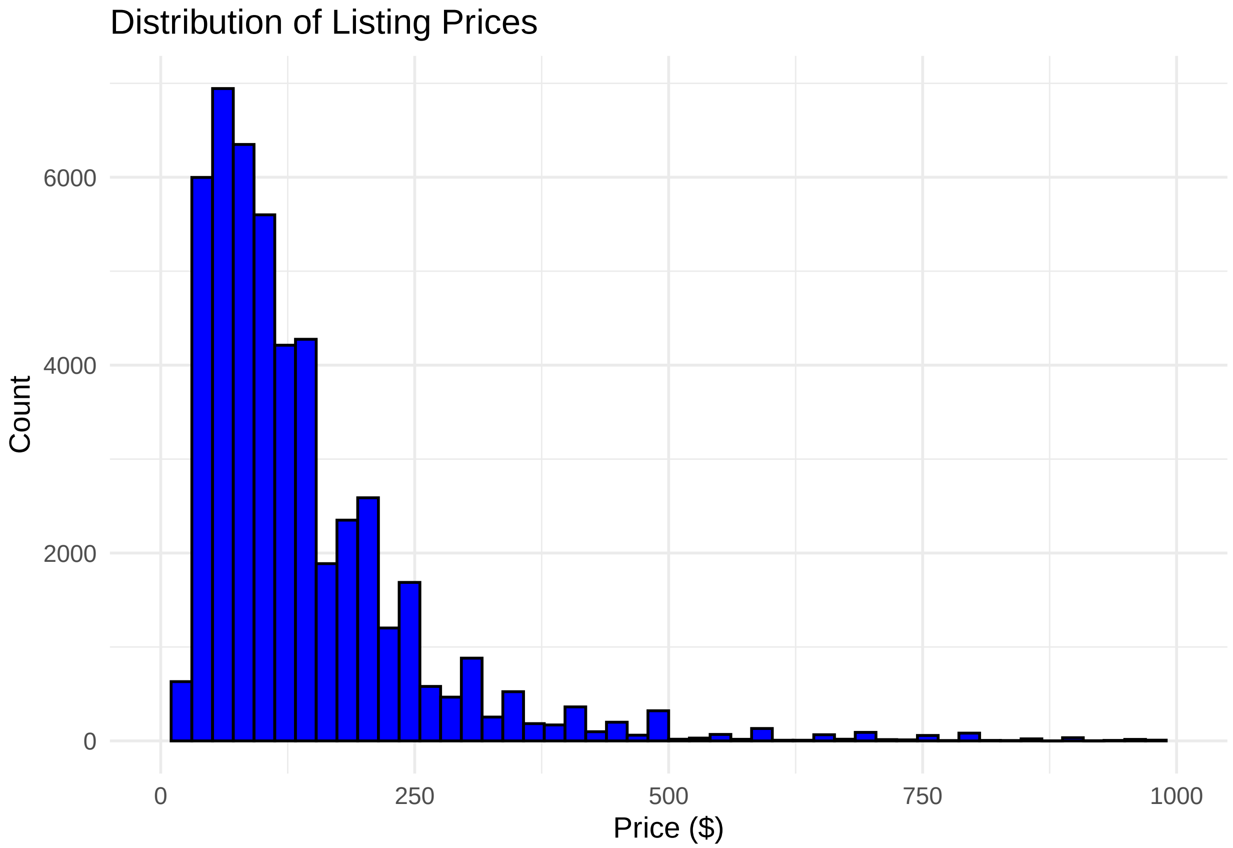
```
##      price          minimum_nights    number_of_reviews availability_365
##  Min.   :   10.0   Min.   :   1.00   Min.   :  0.00    Min.   :  0.0   
##  1st Qu.:   69.0   1st Qu.:   1.00   1st Qu.:  1.00    1st Qu.:  0.0   
##  Median :  106.0   Median :   3.00   Median :  5.00    Median : 45.0   
##  Mean   :  152.8   Mean   :   7.03   Mean   : 23.27    Mean   :112.8   
##  3rd Qu.:  175.0   3rd Qu.:   5.00   3rd Qu.: 24.00    3rd Qu.:227.0   
##  Max.   :10000.0   Max.   :1250.00   Max.   :629.00    Max.   :365.0   
```

```
avg_price_neigh <- airbnb %>%
  group_by(neighbourhood_group) %>%
  summarise(avg_price = mean(price, na.rm = TRUE),
            num_listings = n()) %>%
  arrange(desc(avg_price))

datatable(avg_price_neigh)
```

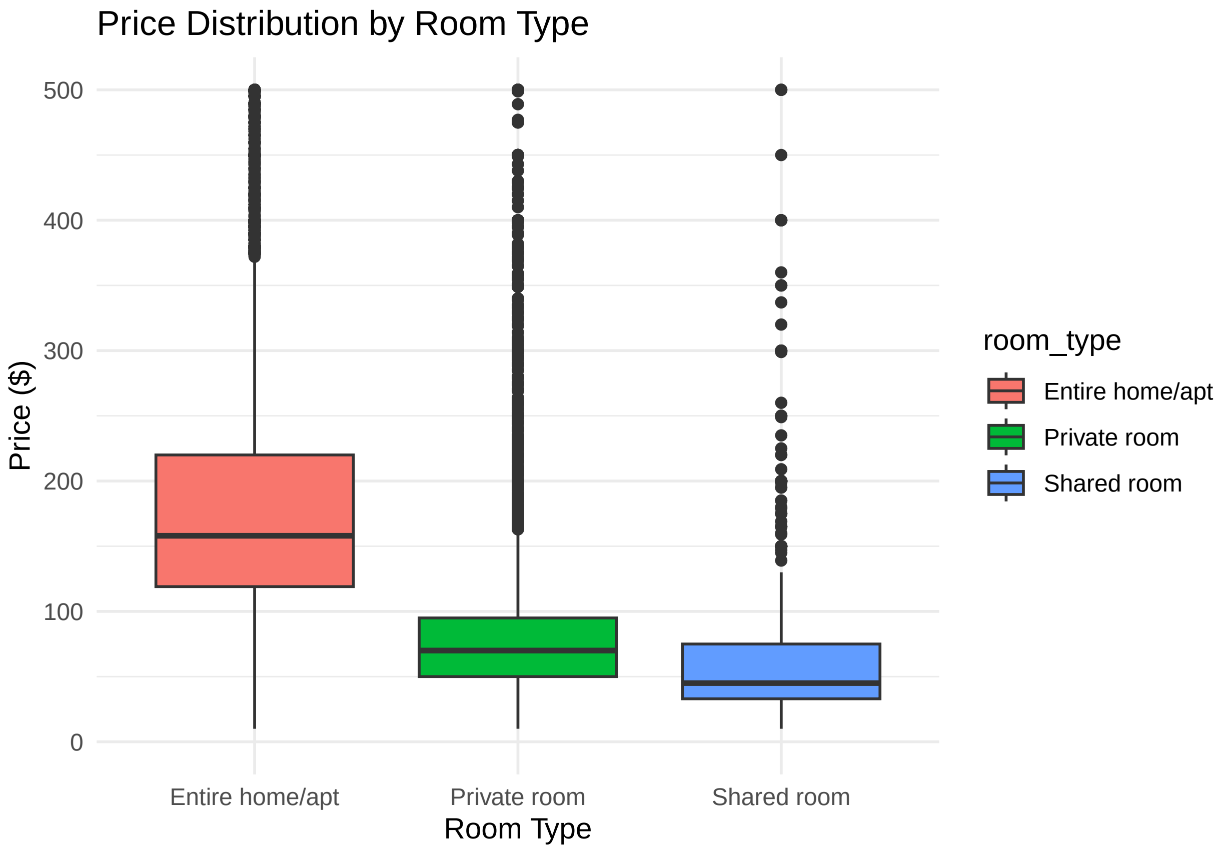
Histogram of prices

```
ggplot(airbnb, aes(x = price)) +
  geom_histogram(bins = 50, fill = "blue", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Listing Prices", x = "Price ($)", y = "Count") +
  xlim(0, 1000)
```



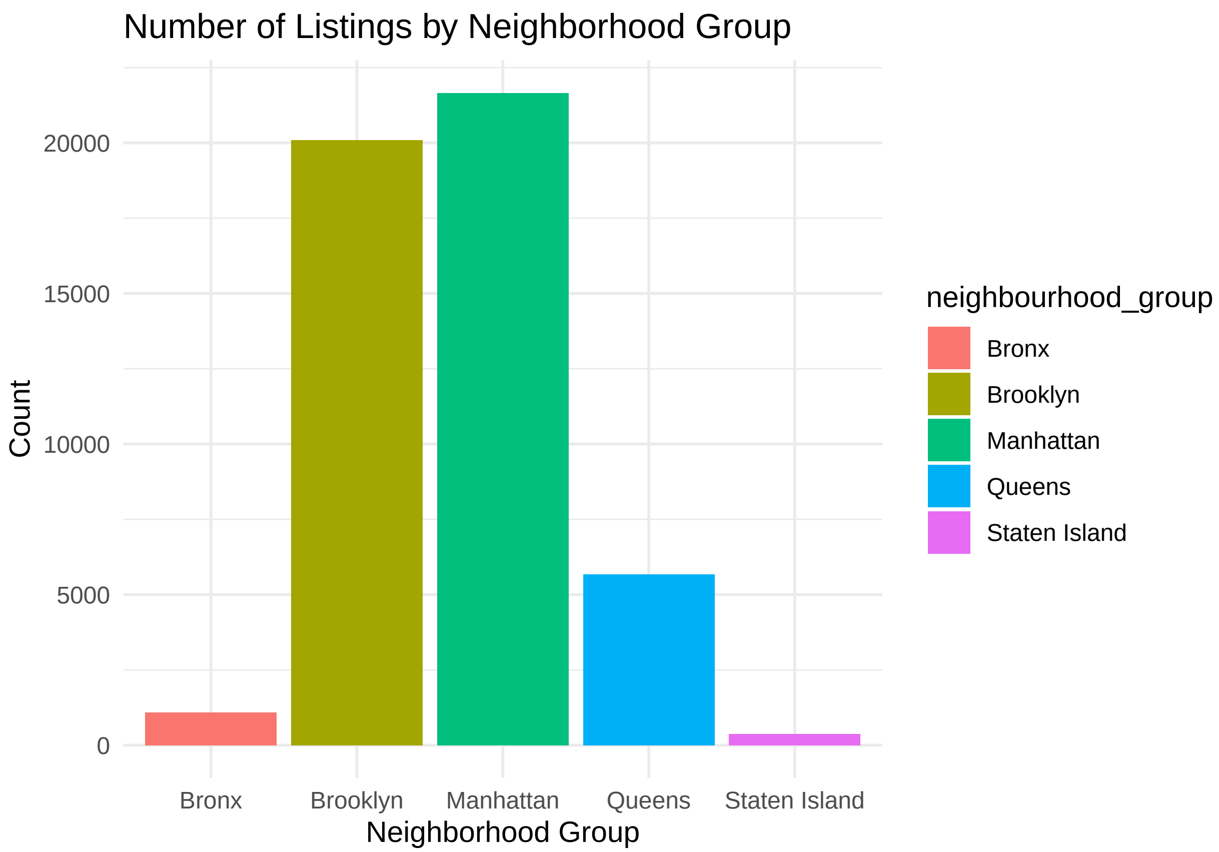
Boxplot of Prices by Room Type

```
ggplot(airbnb, aes(x = room_type, y = price, fill = room_type)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Price Distribution by Room Type", x = "Room Type", y = "Price ($)") +
  ylim(0, 500)
```



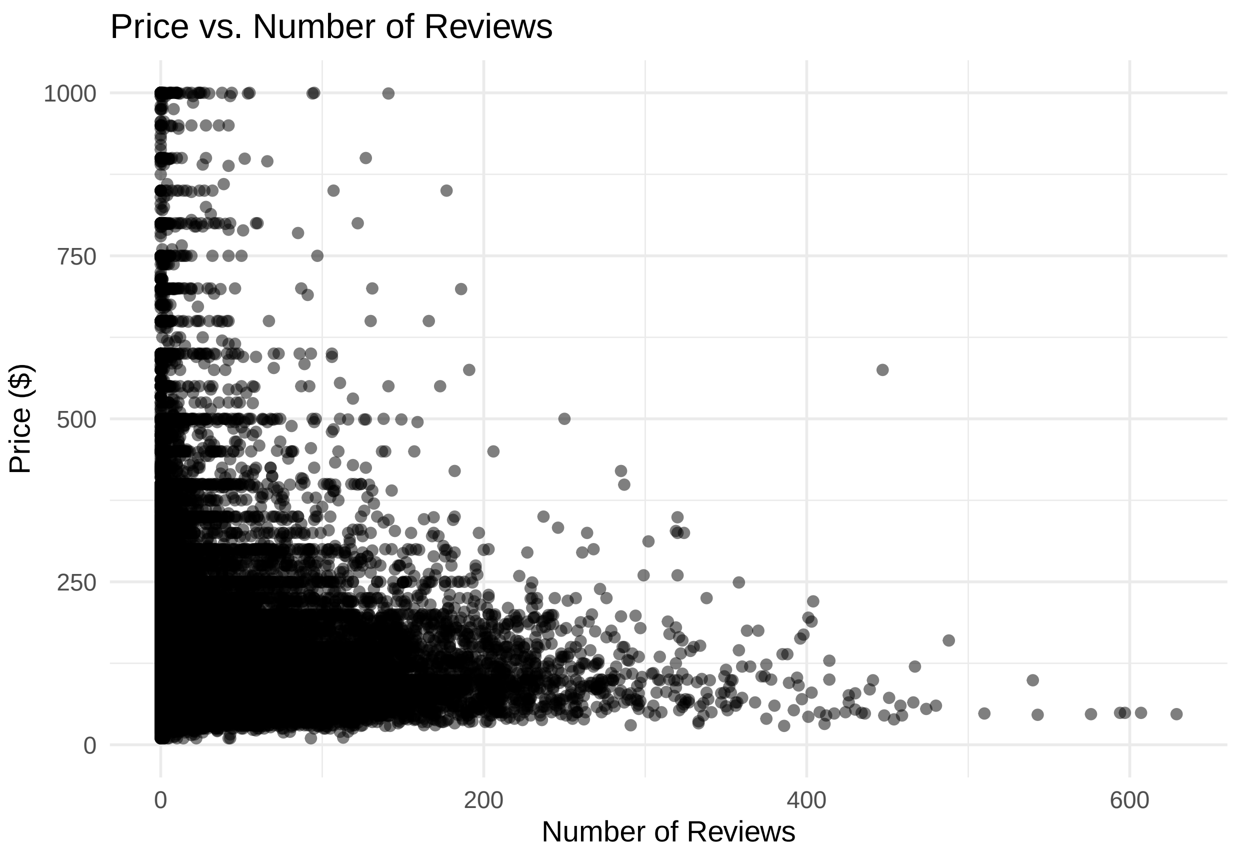
Bar Chart of Listings by Neighborhood Group

```
ggplot(airbnb, aes(x = neighbourhood_group, fill = neighbourhood_group)) +
  geom_bar() +
  theme_minimal() +
  labs(title = "Number of Listings by Neighborhood Group", x = "Neighborhood Group", y = "Count")
```



Scatter Plot: Price vs. Number of Reviews

```r
ggplot(airbnb, aes(x = number_of_reviews, y = price)) +
  geom_point(alpha = 0.5) +
  theme_minimal() +
  labs(title = "Price vs. Number of Reviews", x = "Number of Reviews", y = "Price ($)") +
  ylim(0, 1000)
```



Correlation Analysis

```
# Calculate correlations and round to 3 decimal places
price_reviews_cor <- round(cor(airbnb$price, airbnb$number_of_reviews, use = "complete.obs"), 3)
price_availability_cor <- round(cor(airbnb$price, airbnb$availability_365, use = "complete.obs"), 3)

# Print results
cat("Correlation between price and number of reviews:", price_reviews_cor, "\n")
```

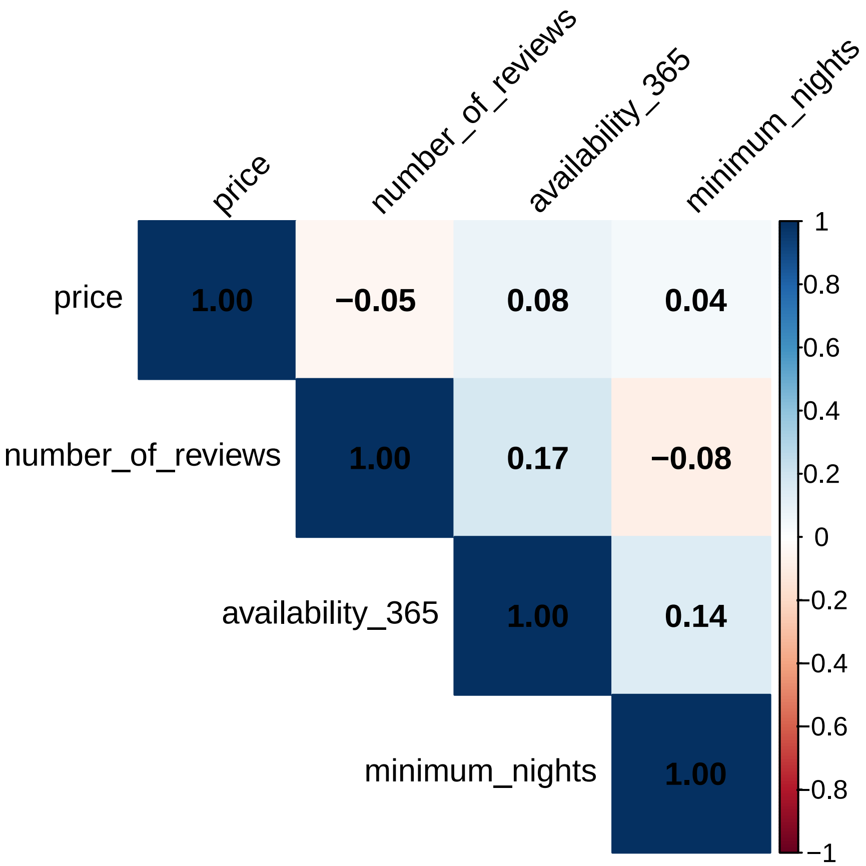
```
## Correlation between price and number of reviews: -0.048
```

```
cat("Correlation between price and availability (365 days):", price_availability_cor, "\n")
```

```
## Correlation between price and availability (365 days): 0.082
```

```
# correlation matrix for numeric variables
cor_matrix <- cor(airbnb %>% select(price, number_of_reviews, availability_365, minimum_nights), use =

# Plot heatmap
corrplot(cor_matrix, method = "color", type = "upper", addCoef.col = "black", tl.col = "black", tl.srt =
```


price
number_of_reviews
availability_365
minimum_nights
price
1.00
−0.05
0.08
0.04
number_of_reviews
1.00
0.17
−0.08
availability_365
1.00
0.14
minimum_nights
1.00
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1